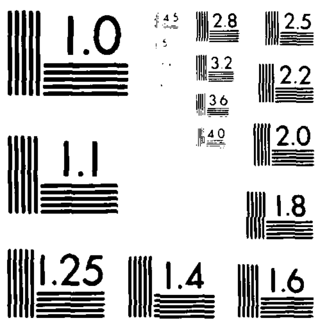
AD-A086 562 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
MEASURES OF LACK OF FIT, AND TRANSFORMATION OF PREDICTORS IN CO--ETC(U)
MAY 80 G E BOX, N R DRAPER DAAG29-75-C-0024
UNCLASSIFIED MRC-TSR-2077 NL

1 OF 1
AD
A086562

END
DATE
FILMED
8-80
DTIC



MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS 1963 A

LEVEL II

ADA086562

MRC Technical Summary Report # 2077

MEASURES OF LACK OF FIT, AND TRANSFORMATION OF PREDICTORS, IN COMPOSITE DESIGNS

G. E. P. Box and N. R. Draper

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

May 1980

(Received February 27, 1980)

Approved for public release
Distribution unlimited

Sponsored by

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

MEASURES OF LACK OF FIT, AND TRANSFORMATION OF PREDICTORS, IN COMPOSITE DESIGNS*

G. E. P. Box and N. R. Draper

Technical Summary Report # 2077
May 1980

ABSTRACT

The central composite designs are investigated in terms of their ability to detect lack of fit and it is shown that measures of non-quadraticity can be obtained in all the k major directions. The geometrical interpretation of this is explained. The relationship of these lack of fit measures to the estimation of power transformations in the predictor variables is also explored.

AMS(MOS) Subject Classification: 62K99

Key Words: Lack of fit, Transformation of predictors, Composite designs

Work Unit Number 4 - Statistics and Probability

*Appeared as Department of Statistics Report #603, University of Wisconsin, Madison.

Sponsored by the United States Army under Contract Nos. DAAG29-75-C-0024 and DAAG29-80-C-0041.

SIGNIFICANCE AND EXPLANATION

A popular and useful class of experimental designs for fitting a second degree equation is the "central composite" class which consists of two-level factorials, plus axial points, plus center points. These are investigated in terms of their ability to detect lack of fit and it is shown that measures of non-quadraticity can be obtained in all the k major directions. The geometrical interpretation of this is explained. The relationship of these lack of fit measures to the estimation of power transformations in the predictor variables is also explored.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the authors of this report.

MEASURES OF LACK OF FIT, AND TRANSFORMATION
OF PREDICTORS, IN COMPOSITE DESIGNS*

G. E. P. Box and N. R. Draper

1. INTRODUCTION

In the course of a response surface study (see, for example, Box, Hunter, Hunter, 1979) we may desire to fit a first (all $\beta_{ij} = 0$) or second order model

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \left\{ \sum_{i=1}^{k} \sum_{j \geq i}^{k} \beta_{ij} x_i x_j \right\} + \varepsilon, \qquad (1.1)$$

relating a response y and k coded predictor variables $x_1, x_2, \ldots, x_k$, where, in practice, both the response and the predictor variables may have been transformed. There are a number of properties which, depending on the investigational circumstances, might be important in choosing a design for fitting such a function (1.1); fourteen were listed by Box and Draper (1975). One characteristic that distinguishes a good response surface design is its efficient use of the available degrees of freedom to achieve the properties desired. In particular, the problem of lack of fit merits special consideration. Whenever we fit a model we must always be concerned with the possibility that some greater degree of complexity may be needed than that allowed for already. We can, of course, employ a design which contains more runs and is suitable for fitting the more complex model, but then similar questions arise concerning <u>that</u> model, and so on. Clearly we cannot guard against all possibilities. A practical compromise is to

*Appeared as Department of Statistics Report #603, University of Wisconsin, Madison.

Sponsored by the United States Army under Contract Nos. DAAG29-75-C-0024 and DAAG29-80-C-0041.

1. employ, at each stage of the experimental iteration, a model which we <u>hope</u> will be adequate but to

2. provide, in the accompanying design, for lack of fit checks that are sensitive to the feared discrepancies, and to

3. use, whenever possible, design arrangements which can be augmented to form larger designs appropriate for fitting and checking the more complex model, if the need for the latter is revealed by the diagnostic checks.

2. MEASURES OF LACK OF FIT

First order designs.

Useful first order designs are the two-level factorials and fractional factorials of resolution three or more which use (respectively) all or some of the 2^k runs $(\pm1,\pm1,\ldots,\pm1)$. As discussed, for example, by Box and Wilson (1951), fractions can be chosen so that checking functions are associated with the residual degrees of freedom containing feared interactions. Moreover, if we limit the choice to designs of resolution at least four, all coefficients in the first order model are rendered free of second order aliases. Alternatively, or in addition, (see De Baun, 1956), by adding n_0 center points to any such design, a contrast between the average response at the center and the average response at the factorial points is made available. This contrast provides an estimate of the sum $(\beta_{11}+\beta_{22}+\ldots+\beta_{kk})$ of the pure quadratic coefficients, if a second order model involving these coefficients were, in fact, needed. It thus provides a check for the need for second order terms in the common situation where we are approximating a (possibly ridgy) maximum or minimum and consequently the β_{ii} are either of the same sign or are near to zero.

Second order designs.

One suitable class of second order designs (see Box and Wilson, 1951) is of the central composite type. A design of this kind for $k = 3$ with n_0

added center points is defined by the columns headed x_1, x_2, and x_3 in Table 1. This design has $14 + n_0$ degrees of freedom (df) available. Ten of these are used for fitting the model.

There are ten possible third order columns, namely those formed by creating entries of the following form (grouped as shown for our convenience in a moment):

$$(x_1^3, x_1x_2^2, x_1x_3^2);\ (x_2^3, x_2x_1^2, x_2x_3^2);\ (x_3^3, x_3x_1^2, x_3x_2^2);\ x_1x_2x_3. \tag{2.1}$$

The last of these, $x_1x_2x_3$, forms a separate column (shown in Table 1) orthogonal to the first ten. The other nine of equation (2.1) form three sets of three as indicated by the parentheses.

Now suppose that the elements of these third order columns are regressed on those of the first ten columns required to fit the second order model, and residuals are then taken to provide the residual vectors x_{111} (from x_1^3), x_{122} (from $x_1x_2^2$), and so on. It is quickly seen that

$$x_{iii} = (1-\alpha^2)x_{ijj}, \qquad i \neq j. \tag{2.2}$$

Thus, the residual vectors are confounded in three sets of three. Furthermore, x_{111}, x_{222} and x_{333} are mutually orthogonal. These vectors, rescaled so that elements of x_{iii} in the cube portion of the design are ±1's, are shown in Table 1.

Consider now a column x_{iii} in relation to Figure 1, which shows the

Table 1. Available Degrees of Freedom in a 3-Factor Central Composite Design

I	x_1	x_2	x_3	x_1^2	x_2^2	x_3^2	x_1x_2	x_1x_3	x_2x_3	x_{111}	x_{222}	x_{333}	$x_1x_2x_3$	Blocks	Pure error	
1	-1	-1	-1	1	1	1	1	1	1	-1	-1	-1	-1	$-\frac{\alpha^2}{4}$	.	.
1	1	-1	-1	1	1	1	-1	-1	1	1	-1	-1	1	$-\frac{\alpha^2}{4}$	.	.
1	-1	1	-1	1	1	1	-1	1	-1	-1	1	-1	1	$-\frac{\alpha^2}{4}$	.	.
1	1	1	-1	1	1	1	1	-1	-1	1	1	-1	-1	$-\frac{\alpha^2}{4}$	.	.
1	-1	-1	1	1	1	1	1	-1	-1	-1	-1	1	1	$-\frac{\alpha^2}{4}$	.	.
1	1	-1	1	1	1	1	-1	1	-1	1	-1	1	-1	$-\frac{\alpha^2}{4}$	.	.
1	-1	1	1	1	1	1	-1	-1	1	-1	1	1	-1	$-\frac{\alpha^2}{4}$	.	.
1	1	1	1	1	1	1	1	1	1	1	1	1	1	$-\frac{\alpha^2}{4}$	.	.
1	$-\alpha$	.	.	α^2	.	.	.	.	.	$\frac{4}{\alpha}$	.	.	.	1	.	.
1	α	.	.	α^2	.	.	.	.	.	$-\frac{4}{\alpha}$	.	.	.	1	.	.
1	.	$-\alpha$	.	.	α^2	.	.	.	.	.	$\frac{4}{\alpha}$	.	.	1	.	.
1	.	α	.	.	α^2	.	.	.	.	.	$-\frac{4}{\alpha}$	.	.	1	.	.
1	.	.	$-\alpha$	.	.	α^2	.	.	.	.	.	$\frac{4}{\alpha}$	.	1	.	.
1	.	.	α	.	.	α^2	.	.	.	.	.	$-\frac{4}{\alpha}$	.	1	.	.
1	.	.	.	.	.	.	.	.	.	.	.	.	.	$\frac{2\alpha^2-6}{n_o}$		
...										.					$\tilde{M}^*$	
1	.	.	.	.	.	.	.	.	.	.	.	.	.	$\frac{2\alpha^2-6}{n_o}$		

*$\tilde{M}$ is an n_o by (n_o-1) nonsingular matrix, all of whose columns sum to zero. The (n_o-1) columns provide any set of pure error comparisons that might be selected.

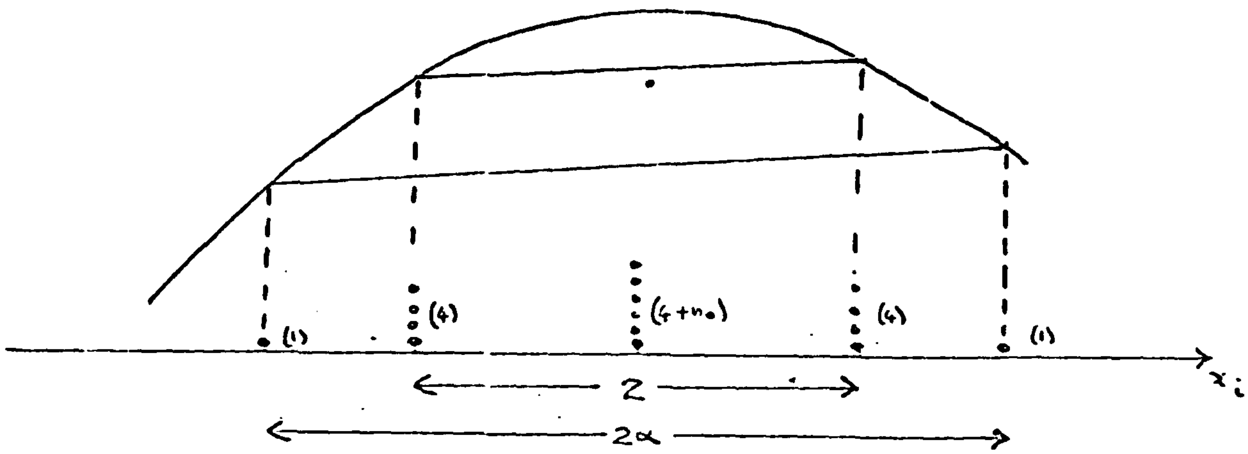


Figure 1. Projection of design points on the x_i axis for a cube plus star plus center points design in three factors; the x_{iii} column provides (apart from a constant) an estimate of the difference between the two slopes shown. The slopes are the same if the true model is quadratic.

projection of the points of the composite design onto the x_i axis. Denoting the average of all the responses at $x_i = -\alpha, -1, 1, \alpha$ by $\bar{y}_{-\alpha}$, $\bar{y}_{-1}$, $\bar{y}_{+1}$, and $\bar{y}_{+\alpha}$, respectively, we see that the contrast associated with x_{iii} is proportional to (in an obvious vector notation)

$$c_i = \frac{1}{8}\{\underset{\sim}{x}'_{iii}\underset{\sim}{y}\} = \frac{\bar{y}_+ - \bar{y}_-}{2} - \frac{y_\alpha - y_{-\alpha}}{2\alpha} \quad . \tag{2.3}$$

Referring to Figure 1, we see that the first of the expressions on the right of equation (2.3) provides an intuitive estimate of the slope of the chord AB, while the second is an estimate of the slope of the chord CD. These chords are parallel if the response is quadratic when $E(c_i) = 0$, but otherwise

$$E(c_i) = (1-\alpha^2)\beta_{iii} + \sum_{j \neq i}^{3} \beta_{ijj} \tag{2.4}$$

and would usually be non-zero. Thus, the contrast c_i provides a natural check on the quadratic model in the x_i direction, for i = 1, 2, and 3.

In general, a composite design is formed of:

(a) A "cube", consisting of a 2^k factorial, or a 2^{k-p} fractional factorial, made up of points of the type $(\pm, \pm 1, \ldots, \pm 1)$, of resolution $R \geq 5$ (Box and Hunter, 1961) replicated $f(\geq 1)$ times. There are thus $n_c = f2^{k-p}$ such points (where p may be zero).

(b) A "star", that is, 2k points $(\pm\alpha,0,0,\ldots,0)$, $(0,\pm\alpha,0,\ldots,0)$,...., $(0,0,0,\ldots,\pm\alpha)$ on the predictor variable axes, replicated r times, so that there are $n_s = 2kr$ points in all.

(c) Center points $(0,0,\ldots,0)$, n_0 in number.

For any particular design, k is given and the values of p, f, α, and n_0 need to be specified.

It is shown in the Appendix that, for any such design, k sets of columns can be isolated with the ith set containing the k columns $x_i x_j^2$, $j = 1,2,\ldots,k$. This ith set is associated with a single vector $\underset{\sim}{x}_{iii}$ which is orthogonal to the $(k+1)(k+2)/2$ columns required for fitting the second degree equation and is also orthogonal to the $(k-1)$ similarly constructed vectors x_{jjj}, $j \neq i$. Adopting a notation similar to that used previously for the $k = 3$ case, we find that the k orthogonal contrasts defined by these vectors are such that, for $i = 1,2,\ldots,k$,

$$c_i = \frac{1}{f2^{k-p}} (\underset{\sim}{x}'_{iii}\underset{\sim}{y}) = \frac{\bar{y}_+ - \bar{y}_-}{2} - \frac{y_\alpha - y_{-\alpha}}{2\alpha} \tag{2.5}$$

and may be interpreted, as before, as checks for parallelism of chords in the k different axial directions.

It may also be shown that, on the assumption of the need for a cubic model,

$$E(c_i) = (1-\alpha^2)\beta_{iii} + \sum_{j \neq i}^{k} \beta_{ijj} \ . \tag{2.6}$$

3. RELATIONSHIP TO TRANSFORMATION OF PREDICTOR VARIABLES.

We mentioned earlier the dilemma in which an investigator invariably finds himself or herself. To run a design appropriate for fitting an unnecessarily elaborate model is wasteful, but to oversimplify might be misleading. In particular, when we hope that a second degree polynomial may be adequate for representation, we shall wish to provide ourselves with sensible checks without going to the extreme of running a complete third order design. We need then to check for what we believe will be the most likely discrepancies (if, indeed, there are any at all). We think that, in many practical situations, the k "pairs of parallel chords" checks which the second order composite design always provides, have a natural appeal.

The problem may also be approached in terms of the possible need for transformation of the predictor variables. It is natural to speculate that, if a second degree equation is not adequate for the original uncoded predictor variables, $\xi_1,\xi_2,\ldots,\xi_k$, say (or their coded versions $x_1,x_2,\ldots,x_k$), it might be adequate if some transformation of the ξ_i were used. The idea obviously relates to the earlier one because, in general, one piece of evidence of the need for transformation would be the lack of chord parallelism in one or more axial directions. What we shall show is that the k contrasts $c_1,c_2,\ldots,c_k$ associated with chord parallelism also contain information provided by the data about the need for transformation of the predictor variables. However, we shall also see that the existence of interactions β_{ij}, $i \neq j$, in the second order model means that the pieces of information provided by the c_i overlap to some extent, that is we cannot relate c_i only to transformation on x_i when $\beta_{ij} \neq 0$.

Suppose, for example, that a composite design is run in the coded variables

$$x_i = (\xi_i - \xi_{io})/S_i \tag{4.1}$$

in a situation where the underlying response function $f(\underset{\sim}{\xi})$ can be approximated by a second degree polynomial in the transformed <u>original</u> variables, namely $\xi_i^{\lambda_i}$. Expanding via a Taylor's series expansion about the values $\underset{\sim}{\lambda}_o = (1,\ldots,1)'$, and cutting off at first order terms provides the approximation

$$\eta = F(\underset{\sim}{\xi}) + \sum_{i=1}^{k} (\lambda_1 - 1) z_i \tag{4.2}$$

where

$$z_i = \left[\frac{\partial F}{\partial \xi_i^{\lambda_i}} \frac{\partial \xi_i^{\lambda_i}}{\partial \lambda_i} \right]_{\underset{\sim}{\lambda} = \underset{\sim}{\lambda}_o} \tag{4.3}$$

$$\left[\frac{\partial F}{\partial \xi_i^{\lambda_i}} \right]_{\underset{\sim}{\lambda} = \underset{\sim}{\lambda}_o} = \beta_i' + 2\beta_{ii}'\xi_i + \sum_{j \neq 1}^{k} \beta_{ij}'\xi_j, \tag{4.4}$$

and

$$\left[\frac{\partial \xi_i^{\lambda_i}}{\partial \lambda_i} \right]_{\underset{\sim}{\lambda} = \underset{\sim}{\lambda}_o} = \xi_1 \ln \xi_i \tag{4.5}$$

$$\simeq -\frac{1}{2}\xi_{io} + \xi_i \ln \xi_{io} + \frac{1}{2}\xi_i^2/\xi_{io}. \tag{4.6}$$

Combining the pieces above, we see that z_i is approximated by a cubic polynomial in ξ, and that the component of z_i orthogonal to the quadratic model is approximated by

$$v_i = \{\beta'_{ii}\xi_{iii} + \frac{1}{2}\sum_{j\neq i} \beta'_{ij}\xi_{iij}\}/\xi_{io}. \tag{4.7}$$

The elements ξ_{iiiu}, ξ_{iiju}, $u = 1,2,\ldots,n$ are obtained by regressing ξ_{iu}^3 and $\xi_{iu}^2\xi_{ju}$ against all the variables $1, \xi_{1u},\ldots,\xi_{ku};\xi_{1u}^2;\ldots,\xi_{ku}^2;\xi_{1u}\xi_{2u},\ldots,\xi_{k-1,u}\xi_{ku}$ which appear in the quadratic portion of the model, and taking residuals. We now rewrite (4.7) as

$$w_i = S_i\{\beta_{ii}x_{iii} + \frac{1}{2}\sum_{j\neq i}^{k} \beta_{ij}x_{iij}\}/\xi_{io}. \tag{4.8}$$

where $\xi_i = \xi_{io} + S_i x_i$, where additional terms of orders ≤ 2 have been absorbed in the quadratic portion of the model, and where x_{iii} and x_{iij} are the standardized forms of ξ_{iii} and ξ_{iij} employed earlier. Also, β_{ij} is the coefficient of the term $x_i x_j$ in the standardized variables, so that $\beta_{ij} = S_i S_j \beta'_{ij}$. We now show how to apply these results to composite designs.

Example. Consider the estimation of the λ's using a composite design, as described in Section 1. The appropriate third order orthogonal polynomials for such a design are found to be

$$x_{iij} = x_i^2 x_j - \theta x_j \tag{4.9}$$

where

$$\theta = \Sigma x_{iu}^2 x_{ju}^2 / \Sigma\, x_{iu}^2 = f/(f+2r\alpha^2), \tag{4.10}$$

and

$$x_{iii} = x_i^3 - \psi x_i \tag{4.11}$$

where

$$\psi = \frac{\Sigma x_{iu}^4}{\Sigma x_{iu}^2} = \frac{f+2r\alpha^4}{f+2r\alpha^2} = \theta(1-\alpha^2) + \alpha^2. \tag{4.12}$$

Each predictor variable in the design can take only five values, namely $(-\alpha,-1,0,1,\alpha)$. The values of the third order polynomials at existing combinations of these values are as follows.

x_i	x_j	x_{iij}	x_{ijj}	x_{iii}
± 1	± 1	$(1-\theta)x_j$	$(1-\theta)x_i$	$(1-\alpha^2)(1-\theta)x_i$
$\pm\alpha$	0	0	$-\theta x_i$	$(1-\alpha^2)(-\theta)x_i$
0	$\pm\alpha$	$-\theta x_j$	0	0
0	0	0	0	0

Note that $x_{iii} = (1-\alpha^2)x_{ijj}$, which exhibits a well-known aliasing feature of composite designs. Thus (4.8) becomes, for this type of design,

$$w_i = S_i\{\beta_{ii}x_{iii} + \tfrac{1}{2}\sum_{j\neq i}^{k}\beta_{ij}(1-\alpha^2)^{-1}x_{jjj}\}/\xi_{io}. \tag{4.13}$$

The possible need for a transformation parameter λ_i for x_i is thus examined as follows:

1. Fit a full second order model to the response data and so obtain estimates b_o, b_i, b_{ii}, b_{ij} of the corresponding β's.

2. Substitute b_{ii} and b_{ij} into (4.13) to obtain on estimated w_i value; call this $\hat{w}_i$. Create the w_i vector by substituting the design points successively in $\hat{w}_i$.

3. Regress the vector of observations $\underset{\sim}{y}$ on to the vectors $\hat{w}_i$, $i = 1,2,\ldots,k$. This (see Eqs. (4.2) through (4.8)) provides estimates $(\hat{\lambda}_i-1)$ of the (λ_i-1).

4. Test, approximately, the reality or otherwise of the (λ_i-1) by the usual t- or F-test, pretending that $\hat{w}_i$ is fixed, i.e., ignoring the fact that estimates b_{ij} and b_{ij} are used in it. Significant $(\hat{\lambda}_i-1)$ indicate the need for transformations of parameter values estimated by the $\hat{\lambda}_i$.

Comments

(a) The $\underset{\sim}{w}_i$ are all linear combinations of $\underset{\sim}{x}_{111}, \underset{\sim}{x}_{222}, \ldots, \underset{\sim}{x}_{kkk}$. Thus, in general, they are not orthogonal to one another, although $\underset{\sim}{x}_{iii}$ and $\underset{\sim}{x}_{jjj}$ are orthogonal for $i \neq j$. This means the $\hat{\lambda}_i$ are correlated with one another, in general. One way of proceeding is to obtain the regression in (3) via a stepwise regression procedure, and to assess and test the $(\hat{\lambda}_i-1)$ in the order they enter the equation.

(b) Eq. (4.13) implies that a transformation in any one predictor variable depends not only on what is happening in the direction of that predictor variable but also depends on what is happening in the directions of the other predictors, in general, if $\beta_{ij} \neq 0$.

(c) If $\beta_{ij} = 0$ all i, j, $i \neq j$ then (4.13) implies that the transformation needed, if any, is directly related to the cubic lack of fit $\underset{\sim}{x}_{iii}$. The regression can be performed directly on $\underset{\sim}{x}_{iii}$ in such a case. This result could be applied to estimate transformations in the canonical directions.

REFERENCES

Box, G.E.P. and Draper, N.R. (1975). Robust designs. Biometrika, 62, 347-352.

Box, G.E.P. and Hunter, J.S. (1961). The 2^{k-p} fractional factorial designs, Part I. Technometrics, 3, 311-351; Part II, 449-458.

Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978). Statistics for Experimenters. New York: Wiley

Box, G.E.P. and Wilson, K.B. (1951). On the experimental attainment of optimum conditions. J. Roy. Statist. Soc. B, 13, 1-45.

DeBaun, R.M. (1956). Block effects in the determination of optimum conditions. Biometrics, 12, 20-22.

ACKNOWLEDGEMENTS

G.E.P. Box was partially sponsored at the Mathematics Research Center, University of Wisconsin, by the United States Army under Contract No. DAAG-29-75-C-0024.

APPENDIX

We here prove the results summarized in Section 2. The full cubic model in k variables $x_1, x_2, \ldots, x_k$ is given by

$$y = \beta_0 + \sum_i \beta_i x_i + \sum_i \beta_i x_i^2$$

$$+ \sum_{i<} \sum_j \beta_{ij} x_i x_j + \sum_i \sum_j \beta_{ijj} x_i x_j^2$$

$$+ \sum_{i<} \sum_{j<} \sum_{\ell} \beta_{ij\ell} x_i x_j x_\ell \qquad \text{(A1)}$$

where all subscripts run 1,2,...,n and the restrictions shown for them are observed where necessary. The form of the $\underset{\sim}{X}$ matrix in the regression model $\underset{\sim}{y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$ when the design consists of f "cubes" (f replicates of a 2^{k-p} factorial design with coordinates of form $(\pm 1, \pm 1, \ldots \pm 1)$) plus r "stars" (r replicates of 2k axial points distance α from the origin) plus n_0 center points is as shown in Table 2. We can denote columns by placing square brackets around the column head; for example $[x_1]$ will denote the x_1 column, and so on. We write

$$F = f2^{k-p}$$

for the number of factorial points.

Table 2. The $\underset{\sim}{X}$ Matrix For a Cubic Model in k Predictors $x_1, x_2, \ldots, x_k$.

x_0	x_1	x_2	…	x_k	x_1^2	x_2^2	…	x_k^2	x_1x_2	…	$x_{k-1}x_k$	x_1^3	$x_1x_2^2$	…	$x_1x_k^2$	…	$x_1x_2x_3$	…
1	$\underset{\sim}{c}_1$	$\underset{\sim}{c}_2$	…	$\underset{\sim}{c}_k$														
1																		
1		$\underline{+}1$'s				1's				$\underline{+}1$'s			$\underline{+}1$'s				$\underline{+}1$'s	
⋮		2^{k-p} design								interaction patterns			each column same pattern as in x_1 column					
1																		
1	$-\alpha$			0	α^2			0				$-\alpha^3$						
1	α			0	α^2			0				α^3						
1	$-\alpha$	r sets permuted down columns		0	α^2			0				$-\alpha^3$						
1	α			0	α^2			0				α^3						
⋮	⋮			⋮	⋮			⋮		$\underset{\sim}{0}$		⋮	$\underset{\sim}{0}$				$\underset{\sim}{0}$	
	0				0							0						
	0				0							0						
	⋮			$-\alpha$	⋮			α^2				⋮						
				α				α^2										
				$-\alpha$				α^2										
1	0			α	0			α^2				0						
1																		
1																		
⋮		$\underset{\sim}{0}$				$\underset{\sim}{0}$				$\underset{\sim}{0}$			$\underset{\sim}{0}$				$\underset{\sim}{0}$	
1																		

All of the cubic columns are orthogonal to all of the other columns with the following exceptions: $[x_i^3]$ is not orthogonal to $[x_i]$, nor to $[x_i x_j^2]$; $[x_i x_j^2]$ is not orthogonal to $[x_i]$, nor to $[x_i^3]$, nor to $[x_i x_\ell^2]$. The first step is to regress the $[x_i^3]$ and $[x_i x_j^2]$ vectors on the $[x_i]$ and take residuals. Because the columns involved are orthogonal to $[x_0]$, no adjustment for means is needed. We denote the "cube portion" of the $[x_i]$ and $[x_i x_j^2]$ vectors by $\underset{\sim}{c}_i$, as indicated in the table. These two sets of residuals are, where the prime denotes transpose,

$$[x_i^3] - \{[x_i]'[x_i^3]/[x_i]'[x_i]\}[x_i]$$

and

$$[x_i x_j^2] - \{[x_i]'[x_i x_j^2]/[x_i]'[x_i]\}[x_i]$$

both of which reduce to multiples $(1-\alpha^2)m$ and m, respectively, where $m = 2\alpha^2 r/(F+2\alpha^2 r)$, of the same vector $[x_{iii}]$, say. For example,

$$[x_{111}]' = [\underset{\sim}{c}_1', d, -d, d, -d \ldots 0, 0, \ldots, 0]',$$

where

$$d = F/(2\alpha r),$$

and where there are r sets of $(d, -d\text{'s})$ in the vector. In general, for $[x_{iii}]'$, $\underset{\sim}{c}_1'$ will be replaced by $\underset{\sim}{c}_i'$ and the position of the $\pm d$'s will correspond to those

of the $\pm\alpha$'s in the corresponding $[x_i]'$ vector. Note that, because $c_i'c_j = 0$, $i \neq j$, it is obvious that $[x_{iii}]$ and $[x_{jjj}]$ are orthogonal.

It follows that the k cubic coefficients β_{iii}, β_{ijj} ($j \neq i, j=1,2,\ldots,k$, otherwise) cannot be estimated individually but only in linear combination, and that the appropriate estimating function for this is

$$\begin{aligned} \ell_{iii} &= [x_{iii}]'\underset{\sim}{y} \\ &= \underset{\sim}{c}_i \underset{\sim}{y}_1 + d(-r\bar{y}_\alpha + r\bar{y}_{-\alpha}) \end{aligned}$$

where $\underset{\sim}{y}_1$ is the portion of $\underset{\sim}{y}$ corresponding to the cube part of the design, and $\bar{y}_\alpha$, $\bar{y}_{-\alpha}$ are, respectively the averages of observations taken at the α and $-\alpha$ axial points on the x_i axis. If we similarly denote by $\bar{y}_+$ and $\bar{y}_-$ the averages of the F/2 observations in $\underset{\sim}{y}_1$ corresponding to 1 and -1 in $\underset{\sim}{c}_i$, respectively, it follows that

$$F^{-1}\ell_{iii} = (\bar{y}_+ - \bar{y}_-)/2 - (\bar{y}_\alpha - \bar{y}_{-\alpha})/(2\alpha).$$

The expected value of $F^{-1}\ell_{iii}$ is

$$E(F^{-1}\ell_{iii}) = F^{-1}[x_{iii}]'\underset{\sim}{X}'\underset{\sim}{\beta}$$

where $\underset{\sim}{X}$ is as in Table 2 and the coefficients of $\underset{\sim}{\beta}$ correspond to the columns in the obvious manner, namely

$$[\beta_0;\beta_1,\beta_2,\ldots,\beta_k;\beta_{11},\beta_{22},\ldots,\beta_{kk};\beta_{12},\ldots,\beta_{k-1,k};\beta_{111},\beta_{122},\ldots,\beta_{1kk};\ldots;\beta_{123},\ldots].$$

Because $[x_{iii}]$ is orthogonal to all columns of $\underset{\sim}{X}$ except the $[x_i^3]$ and $[x_i x_j^2]$ columns, it follows that

$$E(F^{-1}\ell_{iii}) = F^{-1}\{\underset{\sim}{c}_i'\underset{\sim}{c}_i - 2r\alpha^3 d\}\beta_{iii} + F^{-1}\underset{\sim}{c}_i'\underset{\sim}{c}_i \sum_{j\neq i}^{k} \beta_{ijj}$$

$$= (1-\alpha^2)\beta_{iii} + \sum_{j\neq i}^{k} \beta_{ijj}.$$

Thus all the stated results have been established.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

14 MRC-TSR

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 2077, UWIS-DS-80-643	2. GOVT ACCESSION NO.: AD-A086 562	3. RECIPIENT'S CATALOG NUMBER: 9 Technical
4. TITLE (and Subtitle): 6 Measures of Lack of Fit, and Transformation of Predictors, in Composite Designs.		5. TYPE OF REPORT & PERIOD COVERED: Summary Report - no specific reporting period
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): 10 G. E. P. Box and N. R. Draper		8. CONTRACT OR GRANT NUMBER(s): 15 DAAG29-75-C-0024, DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Mathematics Research Center, University of Wisconsin, 610 Walnut Street, Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: Work Unit Number 4 - Statistics and Probability
11. CONTROLLING OFFICE NAME AND ADDRESS: U. S. Army Research Office, P.O. Box 12211, Research Triangle Park, North Carolina 27709		12. REPORT DATE: 11 May 1980
		13. NUMBER OF PAGES: 20
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): 12 24		15. SECURITY CLASS. (of this report): UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Lack of fit, Transformation of predictors, Composite designs

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The central composite designs are investigated in terms of their ability to detect lack of fit and it is shown that measure of non-quadraticity can be obtained in all the k major directions. The geometrical interpretation of this is explained. The relationship of these lack of fit measures to the estimation of power transformations in the predictor variables is also explored.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED 221200

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

LMED

8